5Г

УДК 004.65:004.89

А.Я. Гладун[1], *Ю.В. Рогушина*[2]
[1]Международный научно-учебный центр информационных технологий и систем НАНУ и МОНУ, г. Киев, Украина
[2]Институт программных систем НАНУ, г. Киев, Украина
glanat@yahoo.com, _jjj_@ukr.net

Использование тезауруса предметной области как инструмента представления знаний при повышении эффективности проблемно-ориентированного поиска в Web

Для того, чтобы повысить релевантность поиска информации в Интернете, предлагается использовать знания об информационных потребностях пользователя, отраженных как в описании стоящей перед ним интеллектуальной задачи, так и в его тезаурусе. Это позволяет делать предположения о тематической близости найденных в Web информационных ресурсов той предметной области, которая пертинентна проблеме пользователя.

Введение

В настоящее время главные направления развития информационных технологий (ИТ) связаны с созданием интеллектуальных информационных систем, основанных на извлечении и обработке знаний в соответствующих предметных областях (ПрО). Однако средства, предназначенные для представления знаний, еще недостаточно совершенны, и это часто заставляет людей вновь и вновь искать решения одних и тех же задач.

Одной из наиболее распространенных задач в области ИТ является поиск информационных ресурсов (ИР) в Интернете, локальной сети либо на отдельном компьютере, представленных в различных форматах (текст, графика, мультимедиа и т.д.), для решения той или иной интеллектуальной задачи, стоящей перед пользователем. Конечным результатом поиска может быть:

1) обнаружение ИР (или набора ИР), удовлетворяющего заданным условиям;

2) извлечение из ИР сведений, позволяющих выбрать нужный объект реального мира (примерами таких задач могут быть поиск специалистов, способных выполнить ту или иную работу; подбор экспертов для оценки какого-либо научного или технического проекта, выбор товара в системах е-коммерции; выбор подходящего курса в дистанционном обучении);

3) извлечение из контента ИР правил или закономерностей, позволяющих осуществить логический вывод над имеющимися данными.

По мере развития Интернета обостряется парадокс: вероятность присутствия необходимой информации в глобальном информационном пространстве растет, а вероятность ее нахождения – уменьшается. Это происходит потому, что наполнение Web громадно по объему, очень разнородно, быстро обновляется, плохо поддается структуризации и управлению. Пользователь информационно-поисковых систем (ИПС), как правило, не является специалистом в области ИТ и вследствие этого может при-

менять только наиболее простые и интуитивно понятные средства формулирования своей информационной потребности. Формальный запрос к ИПС – попытка пользователя формализовать свою информационную потребность и, к сожалению, не всегда удачная (либо вследствие низкой выразительной мощности языка создания запросов к ИПС, либо из-за низкой квалификации пользователя). Так, большинство пользователей, обращающихся даже к достаточно простым ИПС Интернета, используют только часть их возможностей – простые запросы, состоящие из 2-3 слов, и не применяют логические операторы и прочие механизмы расширенного поиска [1]. Кроме того, необходимо учитывать, что часть фактов и знаний уже имеются у пользователя, и нет необходимости предоставлять их ему повторно. Следовательно, поисковые механизмы должны оперировать информационными моделями пользователей, задач и информационных ресурсов.

Таким образом, проблема информационного поиска в Web трансформируется в задачу управления знаниями в среде Web.

Сегодня значительные усилия в этом направлении предприняты в рамках проекта Semantic Web. Уже разработан ряд стандартов для представления знаний (OWL), создания метаописаний ИР (RDF) и формирования запросов к ним (SPARQL).

Semantic Web представляет собой лишь надстройку над существующей сетью информационных ресурсов Web, облегчающую обработку информации на семантическом уровне (т.е. ее смысла) поисковыми системами и другими приложениями. Если раньше поисковые машины основное внимание уделяли глубине и способам анализа текстовых данных, то в Semantic Web основными элементами являются информационные объекты и соответствующие им метаданные. Например, информационный объект «Киев» обладает набором метахарактеристик, которые предоставляют данные о его географическом положении, численности населения и т.д.

Постановка задачи

Сегодня основная проблема, возникающая при поиске информации в Интернете, связана с фильтрацией результатов, полученных от различных ИПС, и отбором тех ИР, которые соответствуют реальным информационным потребностям пользователя. Для такого отбора необходимо формализовать представления пользователя об интересующей его проблеме и разработать средства автоматизированного сопоставления этого описания с метаописаниями различных ИР.

1 Онтологии как средство представления знаний

Для успешного решения задачи поиска информации необходимо представить:

– представления пользователя о знаниях той ПрО, которая его интересует, в некоторой форме, пригодной для компьютерной обработки;

– описание проблемы, для которой пользователю нужны эти сведения;

– требования пользователя к тем ИР, котрые могут удовлетворить его информационную потребность.

Важно достигнуть интероперабельности знаний, т.е. того, чтобы знания, сформированные при решении одной задачи, были пригодны при решении других проблем в различных работах ИС. Именно такой формой представления знаний является *онтология* – соглашение об общем использовании понятий, которое содержит средства представления предметных знаний и договоренности о методах соображений. Она может рассматриваться как определенное описание взгляда на мир в конкретной сфере интересов, который состоит из набора терминов и правил использования этих

терминов, которые ограничивают их значение в рамках конкретной ПрО [2]. Онтологии позволяют формализовать знания пользователей о той ПрО, которая их интересует. При этом такие знания становятся доступны другим пользователям и могут применяться в других ИС. Онтологии, описывающие ПрО, могут потом использоваться для решения различных задач, стоящих перед пользователем.

Онтология – это база знаний, описывающая факты, которые предполагаются всегда истинными в рамках определенного сообщества на основе общепринятого значения тезауруса. Она может использоваться как посредник: между пользователем и информационной системой или между членами сообщества, например, между пользователями некоторого корпоративного хранилища данных.

Формальная модель онтологии O представляет собой упорядоченную тройку O=<X,R,F>, где X – конечное множество концептов (понятий, терминов) предметной области, которую представляет онтология O; R – конечное множество отношений между концептами заданной предметной области; F – конечное множество функций интерпретации, заданных на концептах и отношениях онтологии O [3]. Поскольку при обращении к ИПС пользователь должен иметь возможность получить информацию, пертинентную его запросу, то ее поиск должен быть семантически ориентированным. Для этого средства поиска соответствующей запросу информации предлагается организовать на основе онтологии, содержащей описания семантики ресурсов. Онтологии позволяют формально описать конкретные ПрО.

Ряд авторов предлагают методы автоматического и автоматизированного построения онтологий по естественноязыковым документам. В частности, в [4] на основании обзора ряда работ, в которых рассматривается моделирование ПрО в виде концептуальной модели мира, включающей в себя описания базовых понятий, организованных в родовидовые деревья и совокупность связей между ними, предлагается использовать как синонимы понятия *модели* и *онтологии* ПрО. При этом эта концептуальная модель включает в себя описание объектов, понятий и отношений действительности.

Формирование полного семантического представления текста выполняется средствами глобального семантического анализа [5]. Однако задача формирования множеств выделенных в тексте понятий и семантических отношений модели является нетривиальной и на практике реализуема только для узких и четко формализованных ПрО.

При создании онтологий наибольшую сложность представляет формирование множества F, так как этот процесс требует применения специальных навыков из области инженерии знаний и формальной логики. В то же время по трудоемкости основная работа по формированию онтологий приходится на формирование множества X, причем эта работа доступна большинству специалистов произвольной предметной области. Несколько сложнее определить множество отношений R, которые надо использовать для моделирования ПрО, но в большинстве случаев можно использовать стандартные наборы из 10 – 20 базовых отношений («быть частью», «быть подклассом», «являться одинаковым» и т.д.).

В связи с этим представляется целесообразным использовать для моделирования знаний пользователя о ПрО поиска с помощью частного случая онтологии – тезауруса, построение которого относительно проще. До недавнего времени термины «онтология» и «тезаурус» использовались как синонимы, однако теперь в ИТ тезаурус чаще применяют для описания лексики в проекции на семантику, а онтологию – для моделирования семантики и прагматики в проекции на язык представления [6].

Как показывает анализ публикаций, достаточно четко установить взаимоотношение терминов «Тезаурус» и «Онтология» – сложная проблема в связи с расплывчатостью и почти полным сходством их интерпретации. Тезаурус из всего спектра

средств языка отражает только лексику: она задана в знаковом виде и относительно просто поддается систематизации. Тезаурус можно было бы представить как комплекс лингвистических знаний, включающий все составляющие языка от фонетики до риторической структуры текста и законов коммуникации.

2 Тезаурус как средство моделирования ПрО

Обычно тезаурус Т определяют как словарь, содержащий лексические единицы (ЛЕ) с явным указанием семантических связей между ними.

Слово *тезаурус* происходит от греческого *сокровищница, запас*, *клад.* Термин «тезаурус» достаточно древнего происхождения. Впервые его применил в значении, близком сегодняшнему, еще в XIII веке Б. Датини в энциклопедии «Книга о сокровище». Согласно «Современному словарю иностранных слов»: *тезаурус* – 1) словарь, в котором максимально полно представлены все слова языка с исчерпывающим перечнем примеров их употребления в текстах; в полном объеме осуществим лишь для мертвых языков; 2) идеографический словарь, в котором показаны семантические отношения (синонимические, родо-видовые и др.) между лексическими единицами; 3) в информатике – полный систематизированный набор данных о какой-либо области знаний, позволяющий человеку или вычислительной машине в ней ориентироваться. Тезаурус (согласно третьему определению) можно рассматривать как частный случай онтологии. Очевидно, что можно говорить о тезаурусе человечества как о сумме накопленных им знаний. Можно исследовать как тезаурусы отдельных специалистов, так и тезаурусы областей знания.

Впервые тезаурус был использован в связи с вычислительными машинами в 1954 г. А. Мастерман в области машинного перевода. Позднее при помощи тезаурусов устанавливалось соответствие между языком запросов пользователя и документами в информационно-поисковых системах. Но еще в начале 60-х гг. Ю.А. Шрейдер предлагал рассматривать тезаурус как систему знаний, отраженных языком, когда тезаурус становится интересным сам по себе, а не только как вспомогательный инструмент.

Можно рассматривать тезаурус как модель терминологической системы. *Терминологическая система* (ТС) – это сложная динамическая устойчивая система, элементами которой являются отобранные по определенным правилам лексические единицы какого-нибудь естественного языка, а структура изоморфна структуре логических связей между понятиями специальной области знаний и деятельности, а функция состоит в том, чтобы служить знаковой (языковой) моделью этой области знаний и деятельности [5]. Можно говорить о том, что ТС является отображением определенной ПрО.

Тезаурус – это $\mathrm{Ts} = \langle \mathrm{T}, \mathrm{R} \rangle$, где Т – множество терминов, а R – множество отношений между этими терминами. Множества Т и R конечны.

Термин – это слово или словесный комплекс, соотносящийся с понятием определенной организованной области познаний (науки, техники), вступающий в системные отношения с другими словами и словесными комплексами и образующее вместе с ними в любом отдельном случае и в определенное время замкнутую систему, отличающуюся высокой информативностью, однозначностью, точностью и экспрессивной нейтральностью. Слово «*термин»* происходит от латинского «*terminus» – «граница*». Множество терминов тезауруса Т соответствует множеству концептов X онтологии O. Такие свойства терминов и ТС, как системность, устойчивость и регулярность связей, отсутствие экспрессии, установка на объективность описания, делают возможным моделирование ТС с помощью тезаурусов. Классификация понятий ПрО через набор слов, условно синонимичных и образующих класс условной эквивалентности, лежит в основе тезаурусов, используемых для информационного поиска. *База знаний*

(БЗ) – семантическая модель, описывающая структуру ПрО. В состав БЗ ПрО входят онтология ПрО и ее тезаурус. Они используют словарь терминов ПрО, устанавливая отношения между терминами и задавая правила их логического преобразования. Это позволяет отвечать на такие вопросы из этой области, ответы на которые в явном виде не присутствуют в БЗ.

Большинство существующих ИПС имеют развитые средства контекстного поиска документов с учетом морфологической информации о словах. Однако в настоящее время очень незначительное число информационных систем предоставляют возможность тематического поиска, например, поиска с использованием тезауруса. Каждое понятие в тезаурусе может объясняться через набор других понятий, что приводит к появлению семантического поля. Фактически тезаурус пользователя – потребителя информации – это вербализованная совокупность его представлений об исследуемой ПрО (рис.1). Основная цель разработки информационно-поисковых тезаурусов – использование их единиц (дескрипторов) для описания основных тем документов в процессе ручного индексирования.

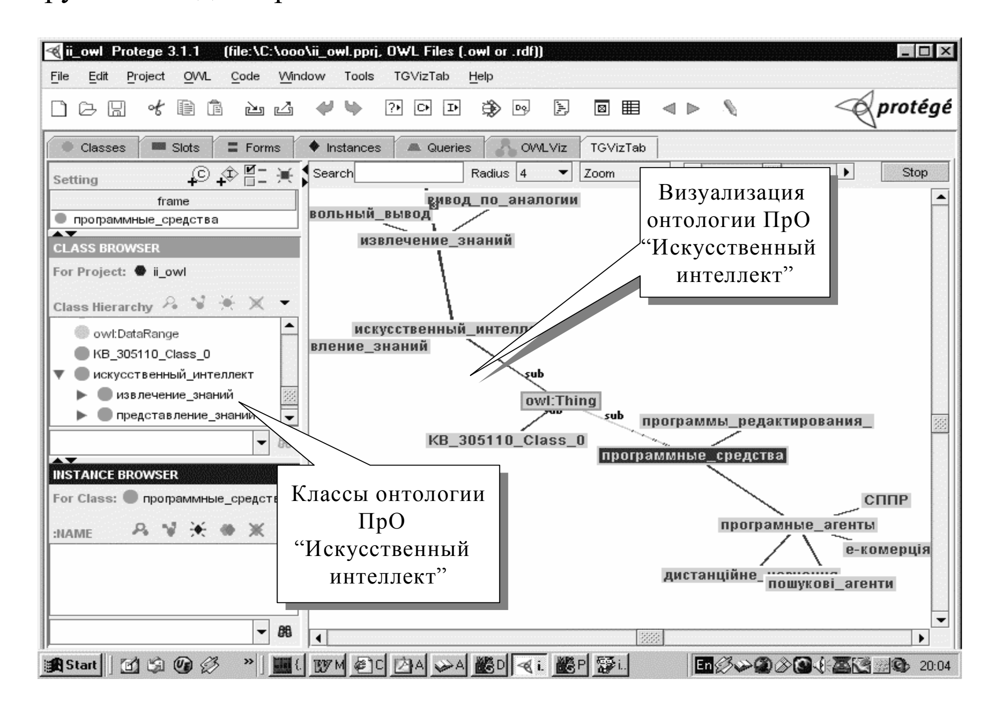


Рисунок 1 – Представление тезауруса ПрО «Искусственный интеллект» в Protégé

Тезаурус может стать эффективным инструментом формирования запросов к универсальным ПМ Internet, для поиска информации в локальной сети, на отдельном компьютере и т.д. Технология полнотекстового поиска является неотъемлемой составляющей таких современных и перспективных ИТ, как: системы управления документами (Document management system, DMS), технологии групповой работы над документами (groupware), технологии поиска в Internet/intranet. Это позволит существенно повысить качество информационного поиска в специализированной тематической области при выполнении следующих условий:

– тезаурус отражает терминологию достаточно узкой научной/предметной области;
– в тезаурусе используются различные семантические отношения;
– тезаурус независим от поисковой машины.

Четко описать терминологию можно при помощи тезауруса с набором сильно дифференцированных семантических отношений [4], [7], т.е. использовать не только

универсальные (например, «род – вид», «часть – целое» и т.д.), но и специфические для конкретной ПрО отношения, несущие значительную смысловую нагрузку.

Кроме непосредственных характеристик тезауруса – количества терминов, количества типов связей и количества реализованных связей, важное значение имеют также их производные – коэффициент связности и количество связных компонент сети. Коэффициент связности показывает, насколько семантическая сеть тезауруса отличается от полного графа (у полного графа любые две вершины смежны, т.е. коэффициент связности равен 1). Для связного графа вычисляется число связности графа – называется наименьшее число вершин, удаление которых приводит к несвязному или одновершинному графу. Анализ этих характеристик позволяет оценить качество тезауруса и сравнивать различные тезаурусы, созданные для одной и той же ПрО.

Основные технологические фазы создания тезауруса:

1. Выделение лексических единиц, т.е. формирование словаря (глоссария) Т.
2. Разработка набора семантических связей.
3. Актуализация связей – установление связей между терминами.

При актуализации семантических связей между терминами тезауруса можно использовать знания экспертов, а также документы, предназначенные как для фиксации структуры знаний ПрО (словари, классификаторы и т.д.), так и отражающие сами знания ПрО (рефераты, статьи, монографии и т.д.).

3 Использование тезаурусов для семантической обработки информации

Чтобы отфильтровать результаты работы внешней ИПС и получить только те ИР, которые пертинентны информационным потребностям пользователя, необходимо предварительно сформировать тезаурус ПрО, интересующей пользователя, и тезаурусы этих ИР, а затем сравнить эти тезаурусы. Построение тезаурусов для информационных ресурсов выполняется программой автоматически на основе лексического анализа соответствующего текста.

Тезаурус используют также для измерения количества информации в ИР на *семантическом* уровне, что позволяет связать семантические свойства информации с возможностью пользователя воспринимать (потреблять, использовать) сообщения, которые поступили по его запросу. Здесь возможны некоторые предельные случаи, например, если количество семантической информации в сообщении равняется нулю, тогда: 1) пользователь вообще не понимает информации; 2) пользователь все знает, а та информация, которая поступила, ему не нужна. Примером первого предельного случая может быть текст на неизвестном пользователю языке, а второго – таблица умножения для студента.

Будем считать, что *тезаурус ПрО* – это совокупность терминов, знакомых пользователю ИПС. Это термины, содержащиеся в ИР, которые были найдены ранее по запросам пользователя и были признаны им относящимися к этой ПрО.

Разработка тезауруса для автоматической оценки семантического количества информации в ИР характеризуется, прежде всего, необходимостью описания значительно большего количества терминов (слов и словосочетаний), встречающихся в текстах данной ПрО. Тезаурус должен включать не только термины, которые представляют важные понятия в текстах данной предметной области, но также охватывать широкий круг более специфических терминов, обнаружение которых в конкретном тексте сделает этот текст релевантным запросу по понятиям более высокого уровня. В результате сопоставления контента ИР с тезаурусом пользователя создается понятийный индекс ИР, в котором указывается, какие дескрипторы тезауруса обнаружены.

4 Алгоритм определения пертинентности ИР информационным потребностям пользователя

4.1 Формирование тезауруса ПрО, интересующей пользователя

На первом этапе пользователь должен создать тезаурус, моделирующий интересующую его ПрО, в котором содержатся основные термины ПрО и связи между ними, и сохранить ее. Для этого можно применить методологию разработки онтологических моделей – стандарт IDEF5 семейства IDEF (www.idef.com/IDEF5.html). Согласно методологии IDEF5, построение тезауруса ПрО состоит из пяти основных действий:

1. *Изучение и систематизирование начальных условий* – цели и контекст разработки тезауруса, определение границ ПрО, интересующей пользователя.

2. *Сбор и накапливание данных* – отбор ИР, относящихся к данной ПрО.

3. *Анализ данных* – изучение отобранных ИР, формирование словаря терминов ПрО, содержащихся в отобранных ИР.

4. *Начальное развитие тезауруса* – установление связей между терминами ПрО (путем формирования пользователем или выбора среди существующих онтологии ПрО, например, с помощью Protégé), из которой затем извлекаются базовые термины ПрО и связи между ними); альтернативным способом построения тезауруса является непосредственный ввод терминов тезауруса пользователем.

5. *Уточнение и утверждение тезауруса* – анализ пользователем полученного тезауруса и его корректирование.

4.2 Формирование тезауруса информационного ресурса

В связи с необходимостью анализа большого количества ИР, мы предлагаем использовать упрощенный алгоритм построения их тезауруса: по полному перечню слов, используемых в ИР, строится словарь терминов, из которого отбрасываются стоп-слова, содержащиеся в специально разработанном пользователем списке. Этот алгоритм применяется только для тех ИР, которые не сопровождаются метаописаниями. В противном случае из метаописаний (в формате RDF или OWL) извлекаются термины тезауруса и связи между ними, которые дополняют построенный по контенту ИР словарь.

4.3 Фильтрация ИР на основе тезаурусов

Алгоритм фильтрации результатов запроса пользователя к внешней ИПС Интернета:

1. Пользователь вводит запрос, идентифицируя свою информационную потребность с помощью ключевых слов.

2. Запрос передается внешней ИПС, от которой получают в соответствии с запросом результаты его выполнения – n ссылок на ИР и их кратких описаний $I = \{Ref_j, D_j\}, j = \overline{1,n}$. Здесь Ref_j – *http*-адрес соответствующего ИР, найденного ИПС, а d_j – информация об этом ИР, которую ИПС предоставляет пользователю в ответ на запрос.

3. Если множество I не пусто, т.е. ИПС найден в ответ на запрос более чем один ИР ($n \geq 1$), то нужно установить порядок, в каком предлагать пользователю сведения о найденных ИР. Тогда для всех ИР из этого множества $I = \{Ref_j, D_j\}, j = \overline{1,n}$ формируются их упрощенные тезаурусы $Ts(ИР_j) = \langle T_{j_j}, \varnothing \rangle, j = \overline{1,n}$ и соответствующие им словари

терминов $T_j = \{t_{j_{lw}}\}, j = \overline{1,n}, w = \overline{1,q_j}$. $t_{j_{lw}}$ – это слова, которые используются в информации о j-м ИР, найденном ИПС, т.е. в $D_j, j = \overline{1,n}$. $q_j, j = \overline{1,n}$ – это количество различных слов, используемых в описании $D_j, j = \overline{1,n}$. Если слова в описании повторяются, то в словаре терминов они фиксируются только один раз.

4. Затем пользователь формирует тезаурус интересующей его ПрО (или указывает на ранее сформированный тезаурус) $Ts_{ПрО}$ и соответствующий ему словарь терминов этой ПрО $T_{ПрО} = \{t_m\}, m = \overline{1,q}$. $T_{ПрО}$ – это множество, состоящее из m терминов, относящихся к интересующей пользователя ПрО. Это множество строится аналогично словарю терминов ИР и обычно формируется как объединение словарей терминов, содержащихся в документах, которые пользователь нашел ранее и посчитал релевантными интересующей его ПрО (как в их контенте, так и в метаописаниях).

5. Производится сравнение $T_{ПрО}$ и $T_j, j = \overline{1,n}$, высчитывается коэффициент их близости $K_j = \sum_{m=1}^{q}\sum_{w=1}^{w_j} f(t_{j_w}, t_m), m = \overline{1,q}, w = \overline{1,w_j}$, где

$$f(t_1, t_2) = \begin{cases} 0, \text{если } t_1 \neq t_2 \\ 1, \text{если } t_1 = t_2 \end{cases}. \qquad (1).$$

Коэффициент (1) представляет собой количество терминов, которые встретились как в тезаурусе ИР, так и в тезаурусе ПрО.

6. Найденные ИР упорядочиваются в зависимости от значений K_j, пользователю предъявляются в первую очередь те ИР, которые имеют наиболее высокий коэффициент близости к ПрО.

При использовании коэффициента (1) возникает следующая проблема: слова, соответствующие одному термину, но являющиеся, например, различными словоформами, синонимами или переводами на различные языки, обрабатываются как разные термины. Поэтому представляется целесообразным использовать онтологию ПрО и выделять группы слов, соответствующих одному термину. Для этого пользователь должен связать элементы словаря терминов терауруса ПрО с одним из терминов онтологии ПрО O = <X,R,F>, т.е. $\forall t_m \in T_{ПрО}, m = \overline{1,q} =$ задать функцию $g(t_m) \in X$. Затем для вычисления коэффициента близости K^O эта функция используется следующим образом:

$$K^O{}_j = \sum_{m,l} f(t_{j_w}, t_m), m = \overline{1,q}, w = \overline{1,w_j}, \text{ где } f(t_1, t_2) = \begin{cases} 0, \text{если } g(t_1) \neq g(t_2) \\ 1, \text{если } g(t_1) = g(t_2) \end{cases}. \qquad (2)$$

Коэффициент (2) представляет собой количество терминов, которые встретились как в тезаурусе ИР, так и в тезаурусе ПрО, и при этом ссылаются на один и тот же термин онтологии ПрО. По сравнению с коэффициентом (1) коэффициент (2) позволяет использовать меньший объем документов для построения тезауруса ПрО, но требует большее время для вычислений.

5 Программная реализация

Предложенные выше методы реализованы в интеллектуальной поисковой системы МАИПС (авт. свидетельство № 32015 и № 32068 от 13.02.2010), которая отвечает ряду требований к приложениям Semantic Web:

1. Для описания ПрО используются онтологии в формате OWL и тезаурусы, для представления которых используется XML.

2. Результаты, получаемые от внешней ИПС, содержат ссылки на ИР, предоставляемые различными провайдерами.

3. МАИПС осуществляет поиск и текстовых, и мультимедийных ИР.

6 Интеллектуальные методы построения тезаурусов ПрО

При создании тезауруса ПрО, которая интересует пользователя ИПС, необходимо явно указать основные понятия ПрО и связи между ними. К сожалению, большинству пользователей достаточно сложно это сделать (даже имея соответствующие знания и применяя их в своей деятельности). На первом этапе формирования тезауруса пользователь может выбрать одно из следующих решений:

– самостоятельно построить с помощью одного из редакторов онтологий онтологическое описание области его информационных интересов;

– найти (например, в Интернете) какую-либо онтологию, представленную на языке OWL, которую описывает ПрО, близкую к области его информационных интересов;

– сформировать множество понятий ПрО, которое содержит наиболее характерные слова и словосочетания, встречающиеся в интересующих его ИР.

Важно определить, какие именно связи между элементами ПрО являются существенными (и их, следовательно, необходимо включить в систему). Не все существенные связи между терминами ПрО могут быть очевидны пользователю, он может воспользоваться для их нахождения *методами индуктивного вывода*.

Существуют независимые подходы к реализации подобных методов: ID3, ACLS, CART и т.д. Наиболее интересным, в связи со спецификой проводимой работы, оказался алгоритм ID3 [8], который специально разработан для извлечения ценной информации из больших объемов слабо структурированных данных. При работе этого алгоритма время вычислений зависит линейно от числа введенных примеров, числа атрибутов, используемых для описания примеров, и числа узлов в строящемся дереве решений. Это качество отличает его от таких известных алгоритмов построения деревьев решений, как INDUCE, SPROUTER, ROTH-P, в которых усилия, требующиеся для решения задачи, резко возрастают вместе со сложностью задачи.

Если методы, подобные МГУА (метод группового учета элементов), предназначены для нахождения закономерностей по набору количественных измерений параметров и полученному по ним результату, то методы, подобные ID3 и его вариациям (C4.5, ID4 и т.д.), предназначены для обобщения опыта экспериментов, параметры и результаты которых описаны через качественные оценки (лингвистические переменные). В большинстве случаев между их значениями невозможно установить даже относительное упорядочение (например, различные симптомы и диагнозы пациентов). К таким задачам относится и задача поиска информации в Интернете. Например, такой существенный параметр ИР, как язык, не может быть описан количественно. ID3 принадлежит к невозрастающим алгоритмам, то есть при добавлении к набору классифицированных примеров новых нужно обрабатывать снова как старые, так и новые примеры.

Предлагается использовать *ID3m* [9] – модификацию ID3 для произвольного (конечного) количества решений. Он также принадлежит к невозрастающим алгоритмам. В данном случае примерами обучающей выборки являются ИР, полученные ранее пользователем в результате запросов к ИПС. Параметрами, по которым они описываются, являются свойства ИР (язык, время создания, размер, формат, право доступа

и т.д.), а также термины тезауруса пользователя. Значения, соответствующие терминам тезауруса, – «Термин отсутствует в ИР», «Термин встречается в ИР редко», «Термин встречается в ИР часто». В качестве результата используется оценка, данная пользователю найденному ИР (качественная оценка, имеющая два и более значений).

На вход алгоритма поступает обучающая выборка H – набор из n классифицированных (получивших одну из возможных оценок) примеров одинаковой размерности. $H = \{h_i\}, i = \overline{1,n}$. Каждый пример из выборки – упорядоченная последовательность значений s атрибутов и результирующего атрибута $h_i = \langle a_1,...,a_s, r\rangle, i = \overline{1,n}$. Значения атрибутов принадлежат конечным множествам: $a_{j_u} \in A_j, j = \overline{1,n}, u = \overline{1,n_j}$, $r_y \in R, y = \overline{1,n_r}$. Если обучающая выборка содержит примеры, в которых все значения атрибутов одинаковы, а решения различны, то введенная информация недостаточна для построения классификационного правила. Если множество примеров пустое, то можно произвольно связать его с любым решением. Если все примеры относятся к одному классу, строится один лист дерева решений, связанный с этим классом. В противном случае необходимо выбрать один из атрибутов и разделить множество атрибутов на подмножества в зависимости от значения этого атрибута и применить алгоритм к каждому из полученных подмножеств.

На каждом шаге работы алгоритма вычисляется, какой атрибут m несет наибольшее количество информации о результате.

$$C_{max} = \max\{C_z, z = \overline{1,s}\} = \max_{z}\{\sum_{i}\sum_{j}\frac{C(a_{z_i} \in A_z, r_j \in R_j)}{d_z}\}, \quad (3)$$

где C(x,y) – количество информации $C(x,y) = \sum_i \sum_j p(x,y) * \lg p(x,y)$, p(x,y) – вероятность одновременного наступления событий x и y, d_m – стоимость получения значения m-го атрибута.

В результате работы алгоритма ID3m формируется дерево решений, в котором каждый лист связан с одним из решений, каждый узел характеризуется именем одного из атрибутов, а выходящие из такого узла ветви – значениями этого атрибута. Такое дерево решений позволяет ИПС по параметрам вновь найденного ИР прогнозировать, как именно оценит его пользователь, и предлагать пользователю в первую очередь те ИР, которые соответствуют его индивидуальным предпочтениям. Так как точные значения вероятностей событий из обучающей выборки неизвестны, то они аппроксимируются на основе рассматриваемого множества примеров.

Выводы

Предложенный в работе подход к поиску информации в Интернете основывается на использовании знаний пользователя о ПрО, характеризующей его информационные потребности. Пользователь может явно указывать интересующие его термины и получать те информационные ресурсы, которые соответствуют его запросу, но содержат также и эти термины. Такой подход ориентирован на пользователя с относительно стабильными информационными потребностями, не являющегося специалистом в области информационных технологий, и позволяет пользователю избежать рутинной работы по фильтрации результатов обращения к ИПС.

Литература

1. Рогушина Ю.В. Использование онтологического описания предметной области для повышения релевантности информационного поиска / Ю.В. Рогушина // Проблемы программирования. – 2003. – № 4. – С. 54-64.
2. Гаврилова Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский. – Спб. : Питер, 2001.
3. Musen M. Domain Ontologies in Software Engineering: Use of Protege with the EON Architecture / M. Musen // Methods of Inform. in Medicine, 1998. – P. 540-550.
4. Андкреев А.М. Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках [Электронный ресурс] / Андкреев А.М., Березкин Д.В. Симаков К.В. – Режим доступа : http://www.inteltec. ru/publish/articles/textan/RCDL2004.shtml.
5. Браславский П.И. Тезаурус как средство описания систем знаний / П.И. Браславский, С.Л. Гольдштейн, Т.Я. Ткаченко // Информационные процессы и системы. – 1997. – № 11. – Серия 2. – С. 16-22.
6. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология [Электронный ресурс] / А.С. Нариньяни. – Режим доступа : http://www.artint.ru/articles/narin/teon.htm.
7. Noy N. The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping / N. Noy, M. Musen // Stanford Medical Informatics. – Stanford Univ., 2003.
8. Quinlan J.R. Discovery rules from large collections of examples: a case study / J.R. Quinlan // Expert Systems in the Microelectronic Age. – Edinburg, 1979. – P. 87-102.
9. Рогушина Ю.В. Применение методов индуктивного вывода для создания прикладных экспертных систем / Ю.В. Рогушина // Разработка и использование информационных технологий в системах управления. – Киев : Ин-т кибернетики им. В.М. Глушкова АН Украины, 1993. – С. 122-128.
10. Гладун А.Я., Онтологии и мультилингвистические тезаурусы как основа семантического поиска информационных ресурсов Интернет / А.Я. Гладун, Ю.В. Рогушина // The Proc. of XII-th Intern. Conf. KDS'2006, (Varna, Bulgaria). – P. 115-121.

А.Я. Гладун, Ю.В. Рогушина
Використання тезауруса предметної області як інструмента представлення знань для підвищення ефективності проблемно-орієнтованого пошуку у Web
Для того щоб підвищити релевантність пошуку інформації у Web, пропонується використовувати знання про інформаційні потреби користувача, відображені як в описі інтелектуальної задачі, що постає перед ним, так і в його тезаурусі. Це дозволяє робити припущення про тематичну близькість знайдених у Web інформаційних ресурсів до тієї предметної області, що пертинентна проблемі користувача.

Anatoly Gladun, Julia Rogushina
Use of the Thesaurus as a Tool of Knowledge Representation in Improving of the Effectiveness of Problem-Based Web Search
In order to improve the relevance of the Web information retrieval the knowledge about users'whsch is information needs reflected in the description of some intelligent problem and thesaurus is proposed to use. It allows to make the assumptions about thematic proximity of Web information resources to the domain pertinent to user's problem.

Статья поступила в редакцию 28.05.2010.